Dublin Business School

QQI

Higher Diploma in Science in Data Analytics

WINTER 2019 EXAMINATIONS

Module Code: **B8IT108**

Module Description: **Data and Web Mining**

Examiner: **Terri Hoare**

Internal Moderator: **Dr Shahram Azizi Sazi**

External Examiner: **Dr Catherine Mulwa**

Date: Wednesday, 4th December 2019
Time: 18:30-20:30

INSTRUCTIONS TO CANDIDATES

Time allowed is 2 hours
Question 1 is mandatory (30 marks)
Answer 2 of the remaining 3 Questions (35 marks each)

All answers should reference literature and case studies as appropriate. Use of scientific calculators only is permitted.

Question 1: MANDATORY **(30 Marks)**

a) Use the AllElectronics Customer Database Table below to build a Naïve Bayes Classifier that can be applied to predict the class label Buys_computer for the unlabelled tuple X given below. All features are nominal.

X = (Age=youth, income=high, student=No, Credit_rating=fair)

Hint: apply the Naïve Bayesian formula

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$where \;\; P(X|C_i) = P(x_1|C_i) * P(x_2|C_i) * \ldots * P(x_n|C_i)$$

$C_1 \;\; defaulted = yes$
$C_2 \;\; defaulted = no$

Electronics Customer DB - Training Tuples - Label Buys_computer

RID	Age	Income	Student	Credit_rating	Buys_computer
1	youth	high	No	fair	**NO**
2	youth	high	No	excellent	**NO**
3	middle-aged	high	No	fair	**YES**
4	senior	medium	No	fair	**YES**
5	senior	low	yes	fair	**YES**
6	senior	low	yes	excellent	**YES**
7	middle-aged	low	yes	excellent	**YES**
8	youth	medium	No	fair	**NO**
9	youth	low	yes	fair	**YES**
10	senior	medium	yes	fair	**YES**
11	youth	medium	yes	excellent	**YES**
12	middle-aged	medium	No	excellent	**YES**
13	middle-aged	high	yes	fair	**YES**
14	senior	medium	No	excellent	**NO**

[10]

b) A classifier is built to identify bowel cancer. The confusion matrix for the validation dataset is given below. Calculate and interpret the related performance measures of accuracy, recall, precision, specificity, and sensitivity.

	Condition positive	Condition negative
Test outcome positive	(TP) = 20	(FP) = 180
Test outcome negative	(FN) = 10	(TN) = 1820

[10]

c) The term "cat" appears 300,000 times in a 10,000,000 million document-sized corpus (i.e. the web) and 12 times in a single 100-word document. Calculate and interpret the TF-IDF score for the word cat for this 100-word document.

[5]

d) Propose and describe a data preparation method for normalising data.

[5]

ANSWER (2 OUT OF 3) QUESTIONS BELOW

Question 2 **(35 Marks)**

a) Select and critique an appropriate algorithm for building a model using supervised learning. Your answer should include either a description or pseudo-code of how the algorithm works, input and output limitations, pros, cons, hyper-parameters that require optimisation and types of suitable applications.

[15]

b) A 10-fold cross validation is performed on four data sets using three machine learning models: Random Forest, Logistic Regression, and a k-Nearest Neighbour. The training error results are presented in the table below.

Data Sets	Random Forest	Logistic Regression	5-NN
Diabetes	27.6% +/- 4.5%	24.4% +/- 4.2%	28.3% +/- 3.0%
Ionosphere	8.8% +/- 5.2%	14.2% +/- 5.8%	15.7% +/- 4.8%
Sonar	33.3% +/- 9.3%	25.9% +/- 8.9%	19.3% +/- 12.6%
Wine	16.1% +/- 1.8%	10.9% +/- 3.4%	16.3% +/- 3.3%

Analyse the results and discuss the use of cross-validation as a standard approach for validating the accuracy of a predictive model.

[8]

c) Ensemble models improve the error rate and reduce the bias of individual models by aggregating (voting between) the predictions of several base models to produce the final model.

Propose four ways for ensuring base models are diverse.

[12]

Question 3 **(35 Marks)**

a) Select and critique an appropriate clustering algorithm for building a model using un-supervised learning. Your answer should include either a description or pseudo-code of how the algorithm works, input and output limitations, pros, cons, hyper-parameters that require optimisation and types of suitable applications.

[20]

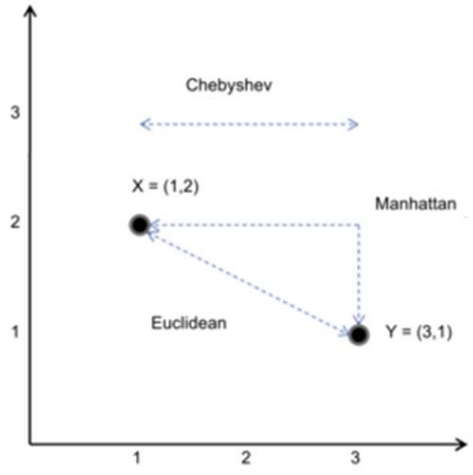
b) Propose a novel use case for a SOM (Self Organising Map).

[6]

c) Minkowski or p-norm distance between two datapoints in n-dimensional space is given by $d = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$.

For the datapoints represented below, calculate

a. Manhatten (taxicab) distance (p=1)
b. Euclidean distance (p=2)
c. Chebyshev distance (p=∞)



[9]

Question 4 **(35 Marks)**

a) Select ONE of the special use cases below and detail the methods, tools, and algorithms that you would use to mine the data.

i. Sentiment analysis of a blog;

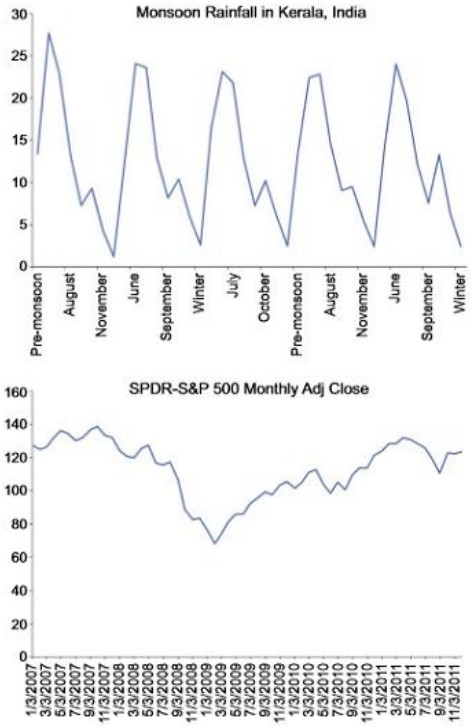
ii. Product demand time series forecasting;

[15]

b) PageRank is an algorithm used by Google Search to rank websites. Critique the measure of network centrality that it uses and contrast with at least two other measures of 'centrality' or power in a network graph.

[10]

c) The figures below show two different time series: annual monsoon precipitation in Southwest India averaged over a five-year period and the adjusted month-end closing prices of the SPDR S&P 500 (SPY) Index over another five-year period. For each time series, propose either a data or model driven forecasting approach motivating the selection of the approach in each case.



[10]

END OF EXAMINATION